ARCHON Augmented: Planning and Web-Enhanced Components

Megan Mou
Stanford University
meganmou@stanford.edu

Sherry Xie
Stanford University
ycxie@stanford.edu

Emily Zhang
Stanford University
emily49@stanford.edu

Andrew Park
Stanford University
aspark@stanford.edu

Abstract

ARCHON is a modular framework which selects and utilizes a variety of inference-time techniques to optimize LLM response generation. However, ARCHON is limited in its number of inference-time components, its lack of web search capabilities, and its lacking performance using relatively smaller open-source models. In this paper, we propose an improved multi-component ARCHON system, using a combination of additional planner, expander, and web search tool call components to improve the ARCHON architecture performance. We evaluate ARCHON and our improved architecture on a new dataset, Humanity's Last Exam, an unsaturated benchmark for frontier academic knowledge and reasoning. We find that even though reasoning remains difficult for ARCHON, our multi-component ARCHON system allows open-source models to rival closed-source performance and that our web search tool call component significantly improves the instruction-following capabilities of ARCHON. We make our code publicly available at: https://github.com/49emily/cs329a-archon/.

1 Introduction

With the rising importance of using inference-time techniques to improve model capabilities, ARCHON proposes using a wide variety of LLMs and inference-time techniques to generate LLM systems more powerful than simply the combination of them (6). However, promising ARCHON performance in existing benchmarks almost all rely on LLMs with about 70B parameters and different models also achieve varied performance on sub-tasks within each query category. As a result, we identified the following potential problems with ARCHON with the hope of improving ARCHON performance especially on smaller, open-source models:

1. ARCHON only contains seven LLM inference time techniques and there is room to explore and test more techniques for a wider variety of tasks.
2. ARCHON does not explore Web Search or tool use capabilities to enhance answer generation.
3. ARCHON pre-defines the category of queries passed into the system without dynamically adjusting its architecture on a query-by-query basis. Queries are often complex and may not fit fully into one category. Dynamic customization of possible ARCHON architecture can lead to more effective and efficient solution for the query at hand.
4. ARCHON explores combinations of 10 SOTA all-source LLMs in its architecture. With the rise of more open-source models and development of more advanced LLM models,

we would love to incorporate more SOTA models into the ARCHON architecture and test ARCHON capabilities in smaller open-source models.

As a result, we set out to answer the question: **Can we improve ARCHON performance by adding more inference-time components and optimizing its architecture to enhance both closed-source and open-source models?** We explored this research question by developing three new inference-time techniques (Planner, Expander, and Web Search Tool Call) and ran ablation studies exploring them with existing ARCHON architecture. We found that expander effectiveness depends strongly on the benchmark, our components can help open-source models rival closed-source model performance with the base ARCHON framework, web-search can significantly improve instruction following, and that reasoning remains a difficult task even with newly added components.

2 Related Work

As mentioned above, we are directly building off of the approach presented in ARCHON, which is a modular framework for selecting, combining, and stacking layersof inference-time techniques to build optimized LLM systems for specific types of benchmarks. In particular, ARCHON already builds off of works such as Mixture-of-Agents (MOA) and LLM-Blender that also fall under the umbrella of multiple-LLM inference-time architectures, but are limited in exploration scope and not as generalizable beyond certain tasks (9) (4).

ARCHON's contributions include defining seven different types of LLM components (Generator, Fuser, Critic, Ranker, Verifier, Unit Test Generator, and Unit Test Evaluator) that are then combined in layers to form an entire chained system. Then, they conduct inference-time architecture search (ITAS) to narrow the search space for possible search hyperparameters, and ultimately develop both general-purpose and task-specific ARCHON architectures that perform better on different respective evaluation datasets.

In their limitations section, the ARCHON authors state that the "addition of new techniques is a promising avenue for future research". Therefore, we were motivated to develop at least 1-2 more unique types of components and experiment further with different architectures in our project, in order to make ARCHON more robust and query-adaptive (beyond the two general categories of instruction-following / reasoning and coding established in the paper).

Specifically, inspired by prior work on planning and multi-step reasoning such as *ReAct: Synergizing Reasoning and Acting in Language Models* that help LLMs generate reasoning traces and interact with external sources, we knew we wanted to incorporate a layer dedicated exclusively to planning (10). Additionally, *Toolformer: Language Models Can Teach Themselves to Use Tools* showed us how powerful tool-calling can be for enhancing model performance on a variety of queries at inference time (7).

3 Methodology

Our key methodology centers on extending the original ARCHON architecture by designing and integrating three novel inference-time components: the **Planner**, the **Expander**, and the **Web Search Tool Call**. These modules are designed to enhance reasoning, context-awareness, and instruction-following capabilities. An overview of our enhanced architecture is shown in Figure 1, and prompts for each component are included in Appendix A.1. We build on top of the existing ARCHON code repository (1).

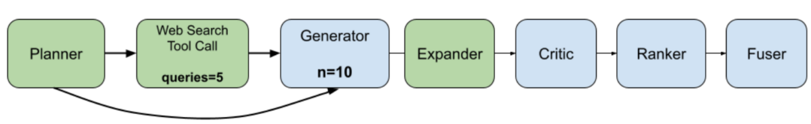

Figure 1: Enhanced ARCHON architecture with Planner, Expander, and Web Search Tool Call

The three novel components introduced are as follows:

1. **Planner:** Positioned at the start of the pipeline, the Planner component rephrases the user prompt and generates a clear step-by-step execution plan. This plan is prepended to subsequent module inputs, ensuring consistent user intent is maintained throughout. This planner component can not only clearly summarize and rephrase the user prompt and capture the original user intent, but can also add even more context to the user query to help tackle hard reasoning questions we evaluate ARCHON on different benchmarks. As mentioned above in prior work, We were inspired by many of the papers from class discussing how to improve LLM planning (10).
2. **Expander:** Following each Generator in the ARCHON architecture, the Expander module enhances the candidate responses by adding context, elaboration, or clarification. This augmentation helps downstream modules—such as Critics or Rankers—better assess the quality of each candidate response.
3. **Web Search Tool Call:** To address ARCHON's lack of external grounding, we introduce a Web Search module placed after the Planner but before the Generator. Inspired by our Homework 2 assignment, this module first generates q queries based on the reasoning trace from the Planner, and then retrieves up-to-date and relevant contextual information by calling the Google Search API, fetching the top 10 webpage results from each query. This up-to-date and relevant contextual information is used as context for the subsequent generation module.

We also considered implementing other components such as a rewriter, summarizer, or debugger, but found these other modules to be either too similar to existing ARCHON components (especially the verifier) and/or too costly to justify a marginal improvement. Overall, we hypothesized that including an explicit planner at the beginning of the entire architecture would have the greatest impact on improving performance, after learning about how crucial, yet difficult, reasoning still is for LLMs (8).

4 Experiments

4.1 Datasets

We focused on two different datasets when running our experiments: **AlpacaEval** and **Humanity's Last Exam**.

AlpacaEval is an instruction-following benchmark designed to test LLMs on real-world tasks. It evaluates models using automated pairwise comparisons against top-performing LLMs like GPT-4 and Claude 3.5. Unlike factual benchmarks, AlpacaEval measures response quality, clarity, and adherence to instructions (3). We use it to ensure ARCHON's inference-time optimizations improve usability and human alignment while maintaining accuracy. Since **AlpacaEval** was used as a dataset in the original ARCHON paper, we think running our experiments against this dataset will easily help us identify the benefits and drawbacks of our methodology.

In addition, we also used the **Humanity's Last Exam** dataset to run our experiments. This is a multi-modal benchmark at the frontier of human knowledge, designed to be the final closed-ended academic benchmark of its kind with broad subject coverage. It contains 2,700 challenging questions across over a hundred subjects, and is one of the prominent benchmarks that has been unsaturated by frontier LLMs (5). We use a subset of HLE (500 questions, limited by compute) to evaluate whether Archon's improvements enhance deep reasoning, problem-solving, and model calibration at the highest difficulty level. We chose this benchmark because HLE is a relatively unsaturated dataset which can easily expose strongsuits and inefficiencies of our improved ARCHON architecture.

4.2 Ablation Experiments

To isolate the impact of each added component, we conducted ablation studies comparing the full multi-component ARCHON system with versions missing one or more of our proposed modules. These studies aimed to answer two core questions: (1) which components contribute the most to performance across benchmarks, and (2) how do different module combinations affect open-source vs. closed-source model capabilities. We ran controlled experiments across both AlpacaEval and HLE to examine module effectiveness under different task demands—specifically instruction-following versus deep reasoning.

4.3 Configuration Details

We evaluated our improved ARCHON architecture under two major configurations: an open-source configuration and a closed-source configuration. Both setups used different sets of LLMs but shared the same experimental protocols, benchmarks, and component layering.

Closed-Source Configuration: This setup uses OpenAI's GPT-4o for both generation and ranking, with Qwen2-72B-Instruct acting as the critic and fuser. This allows us to leverage the strengths of proprietary models while evaluating the marginal benefits of our added inference-time modules.

Open-Source Configuration: For open-source testing, we used an ensemble of models including Qwen2-72B-Instruct, DeepSeek R1 Distill, WizardLM-2 8x22B, QwQ-32B, LLaMA-3 70B Chat, and Mixtral 8x22B. These models collaboratively handled generation, while Qwen2-72B-Instruct was used across critic, ranker, and fuser layers.

Exact module-by-module architecture for both configurations is detailed in Appendix A.3.

5 Results

We found that on Humanity's Last Exam, the best performance at 4.2% was achieved with a closed-source configuration of Planner + Expander + ARCHON using GPT-4o models. The open-soure configuration of the same setup had an accuracy of 3.8%. These results are slightly better than zero-shot GPT-4o, which achieves a 3.1% accuracy. We found that using a Tool Call module actually decreased performance, since the web results were oftentimes not applicable to the question or included too much miscellaneous text on the webpage, and confused the models. The complete results can be found in Table 2 and Figure 4.

On AlpacaEval, we saw significant improvement through using all of our new modules, which boosted both open-source and closed-source performance. Using Planner + Tool Call + Expander + ARCHON boosted the length-controlled win rate of our open-source config to 73.92% and that of our closed-source config to 73.40%. Most notably, adding our modules, especially the Planner, helped our open-source performance, which started more than 10 points lower than closed-source on base ARCHON, rival and even exceed the closed-source performance. The complete results on AlpacaEval can be found in Table 3 and Figure 5.

Architecture	**Closed Source Models**	**Open Source Models**
ARCHON	3.3% (15 / 453)	3.3% (15 / 453)
Planner + ARCHON	3.3% (15 / 453)	2.6% (12 / 453)
Planner + Expander + ARCHON	**4.2% (19 / 453)**	3.8% (17 / 453)
Planner + Tool Call + Expander + ARCHON	3.3% (15 / 453)	2.9% (13 / 453)

Figure 2: Table of HLE Results

Architecture	**Closed Source Models**	**Mixed Source Models**
ARCHON	67.73%	57.01%
Planner + ARCHON	69.62%	67.06%
Planner + Expander + ARCHON	66.01%	68.30%
Planner + Tool Call + Expander + ARCHON	73.40%	**73.92%**

Figure 3: Table of Length-Controlled AlpacaEval Results

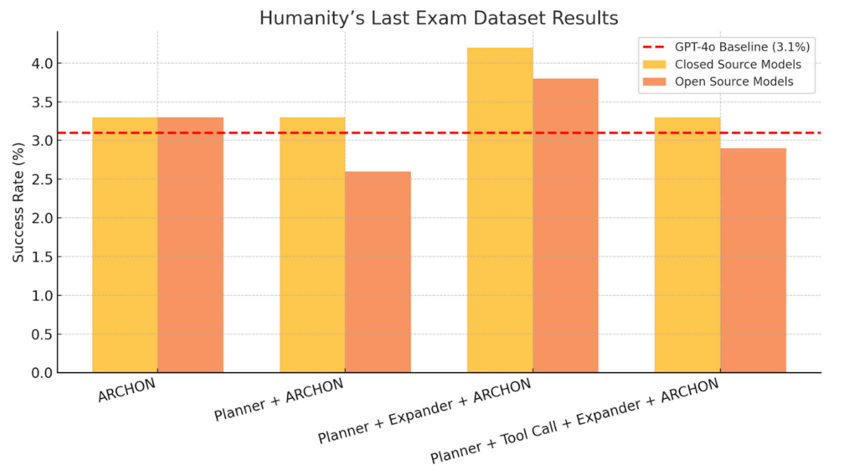

Figure 4: Performance on Humanity's Last Exam

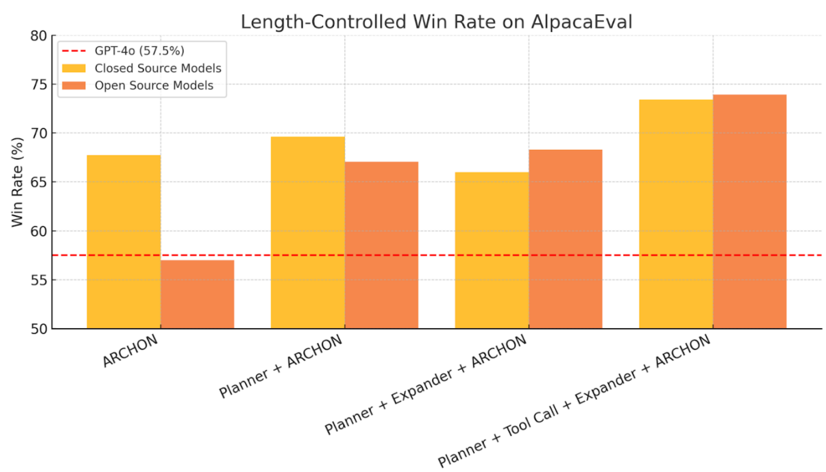

Figure 5: Performance on AlpacaEval

6 Conclusion and Discussion

Our investigation into improving ARCHON through additional inference-time components reveals several key insights. By integrating new modules such as the Planner, Expander, and Web Search Tool Call, we observed notable improvements in model performance across different benchmarks, with varying degrees of effectiveness.

First, our experiments on the length-controlled AlpacaEval dataset indicate that while the Expander module generally improves response quality, its impact is limited due to the dataset's preference for concise answers. However, the Planner module consistently aids in structuring responses more effectively, improving overall performance. Additionally, the Tool Call module significantly enhanced performance, increasing AlpacaEval scores to above 70%, outperforming all modules in the original ARCHON architecture.

Second, we found that incorporating additional inference-time techniques allowed open-source models to rival the performance of closed-source models. Specifically, the Planner and Tool Call modules enabled open-source models to match closed-source performance in instruction-following and information retrieval tasks while maintaining computational efficiency.

Third, our results suggest that the effectiveness of specific modules depends heavily on the benchmark. The Expander module, though less impactful in AlpacaEval, proved beneficial in the Humanity's Last Exam (HLE) dataset by enhancing reasoning capabilities. This highlights the importance of tailoring inference-time techniques to specific tasks and evaluation criteria.

Furthermore, we observed that web search capabilities substantially improve instruction-following accuracy on AlpacaEval. The Tool Call module played a crucial role in grounding responses with external knowledge, leading to more accurate and contextually relevant responses. Yet, web search capabilities were less effective on HLE, requiring more reasoning capabilities than information retrieval. However, the fact that OpenAI's Deep Research (2) scored a new high of 26.6% on HLE suggests that there is still a benefit in external tool calls like web search and room to experiment in using iterative planning/web search loops.

Finally, despite our improvements, reasoning remains a significant challenge. While our efforts with Expander improved performance on HLE, these were not sufficient to overcome reasoning limitations. This suggests that future work should focus on more sophisticated reasoning techniques, such as hierarchical planning or multi-step verification mechanisms.

In summary, our study demonstrates that enhancing ARCHON with additional inference-time components can lead to meaningful performance improvements. However, the success of these techniques depends on the nature of the task, the benchmark used, and the adaptability of the architecture. Future work will focus on a Query-Adaptive Planner to dynamically tailor the architecture per prompt, explore additional tool calls for improved knowledge access, implement early stopping to balance cost and quality, assess whether these modules can offset smaller model sizes without sacrificing accuracy, and investigate additional reasoning-enhancement strategies by leveraging process reward models or other frameworks.

References

[1] Scalingintelligence/archon, 2024. URL: https://github.com/ScalingIntelligence/Archon.

[2] Openai: Introducing deep research, Feb 2025. URL: https://openai.com/index/introducing-deep-research.

[3] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators, 2025. URL: https://arxiv.org/abs/2404.04475, arXiv:2404.04475.

[4] Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion, 2023. URL: https://arxiv.org/abs/2306.02561, arXiv:2306.02561.

[5] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Tung Nguyen, Daron Anderson, Imad Ali Shah, Mikhail Doroshenko, Alun Cennyth Stokes, Mobeen Mahmood, Jaeho Lee, Oleksandr Pokutnyi, Oleg Iskra, Jessica P. Wang, Robert Gerbicz, John-Clark Levin, Serguei Popov, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Mstyslav Kazakov, Geoff Galgon, Johannes Schmitt, Alvaro Sanchez, Yongki Lee, Will Yeadon, Scott Sauers, Marc Roth, Chidozie Agu, Søren Riis, Fabian Giska, Saiteja Utpala, Antrell Cheatom, Zachary Giboney, Gashaw M. Goshu, Sarah-Jane Crowson, Mohinder Maheshbhai Naiya, Noah Burns, Lennart Finke, Zerui Cheng, Hyunwoo Park, Francesco Fournier-Facio, Jennifer Zampese, John B. Wydallis, Ryan G. Hoerr, Mark Nandor, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Jungbae Nam, Edwin Taylor, Jun Jin, Gautier Abou Loume, Hangrui Cao, Alexis C Garretson, Damien Sileo, Qiuyu Ren, Doru Cojoc, Pavel Arkhipov, Usman Qazi, Aras Bacho, Lianghui Li, Sumeet Motwani, Christian Schroeder de Witt, Alexei Kopylov, Johannes Veith, Eric Singer, Paolo Rissone, Jaehyeok Jin, Jack Wei Lun Shi, Chris G. Willcocks, Ameya Prabhu, Longke Tang, Kevin Zhou, Emily de Oliveira Santos, Andrey Pupasov Maksimov, Edward Vendrow, Kengo Zenitani, Joshua Robinson, Aleksandar Mikov, Julien Guillod, Yuqi Li, Ben Pageler, Joshua Vendrow, Vladyslav Kuchkin, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Andrew Gritsevskiy,

Dakotah Martinez, Nick Crispino, Dimitri Zvonkine, Natanael Wildner Fraga, Saeed Soori, Ori Press, Henry Tang, Julian Salazar, Sean R. Green, Lina Brüssel, Moon Twayana, Aymeric Dieuleveut, T. Ryan Rogers, Wenjin Zhang, Ross Finocchio, Bikun Li, Jinzhou Yang, Arun Rao, Gabriel Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Nate Stambaugh, Subrata Mishra, Ariel Ghislain Kemogne Kamdoum, Tad Hogg, Alvin Jin, Carlo Bosio, Gongbo Sun, Brian P Coppola, Haline Heidinger, Rafael Sayous, Stefan Ivanov, Joseph M Cavanagh, Jiawei Shen, Joseph Marvin Imperial, Philippe Schwaller, Shaipranesh Senthilkuma, Andres M Bran, Andres Algaba, Brecht Verbeken, Kelsey Van den Houte, Lynn Van Der Sypt, David Noever, Lisa Schut, Ilia Sucholutsky, Evgenii Zheltonozhskii, Qiaochu Yuan, Derek Lim, Richard Stanley, Shankar Sivarajan, Tong Yang, John Maar, Julian Wykowski, Martí Oller, Jennifer Sandlin, Anmol Sahu, Cesare Giulio Ardito, Yuzheng Hu, Felipe Meneguitti Dias, Tobias Kreiman, Kaivalya Rawal, Tobias Garcia Vilchis, Yuexuan Zu, Martin Lackner, James Koppel, Jeremy Nguyen, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene, Sergey Ivanov, Rafał Poświata, Chenguang Wang, Daofeng Li, Donato Crisostomi, Ali Dehghan, Andrea Achilleos, John Arnold Ambay, Benjamin Myklebust, Archan Sen, David Perrella, Nurdin Kaparov, Mark H Inlow, Allen Zang, Kalyan Ramakrishnan, Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster, Daniel Munro, Linh Ho, Dan Bar Hava, Aleksey Kuchkin, Robert Lauff, David Holmes, Frank Sommerhage, Anji Zhang, Richard Moat, Keith Schneider, Daniel Pyda, Zakayo Kazibwe, Mukhwinder Singh, Don Clarke, Dae Hyun Kim, Sara Fish, Veit Elser, Victor Efren Guadarrama Vilchis, Immo Klose, Christoph Demian, Ujjwala Anantheswaran, Adam Zweiger, Guglielmo Albani, Jeffery Li, Nicolas Daans, Maksim Radionov, Václav Rozhoň, Vincent Ginis, Ziqiao Ma, Christian Stump, Jacob Platnick, Volodymyr Nevirkovets, Luke Basler, Marco Piccardo, Niv Cohen, Virendra Singh, Josef Tkadlec, Paul Rosu, Alan Goldfarb, Piotr Padlewski, Stanislaw Barzowski, Kyle Montgomery, Aline Menezes, Arkil Patel, Zixuan Wang, Jamie Tucker-Foltz, Jack Stade, Declan Grabb, Tom Goertzen, Fereshteh Kazemi, Jeremiah Milbauer, Abhishek Shukla, Hossam Elgnainy, Yan Carlos Leyva Labrador, Hao He, Ling Zhang, Alan Givré, Hew Wolff, Gözdenur Demir, Muhammad Fayez Aziz, Younesse Kaddar, Ivar Ängquist, Yanxu Chen, Elliott Thornley, Robin Zhang, Jiayi Pan, Antonio Terpin, Niklas Muennighoff, Hailey Schoelkopf, Eric Zheng, Avishy Carmi, Jainam Shah, Ethan D. L. Brown, Kelin Zhu, Max Bartolo, Richard Wheeler, Andrew Ho, Shaul Barkan, Jiaqi Wang, Martin Stehberger, Egor Kretov, Peter Bradshaw, JP Heimonen, Kaustubh Sridhar, Zaki Hossain, Ido Akov, Yury Makarychev, Joanna Tam, Hieu Hoang, David M. Cunningham, Vladimir Goryachev, Demosthenes Patramanis, Michael Krause, Andrew Redenti, David Aldous, Jesyin Lai, Shannon Coleman, Jiangnan Xu, Sangwon Lee, Ilias Magoulas, Sandy Zhao, Ning Tang, Michael K. Cohen, Micah Carroll, Orr Paradise, Jan Hendrik Kirchner, Stefan Steinerberger, Maksym Ovchynnikov, Jason O. Matos, Adithya Shenoy, Michael Wang, Yuzhou Nie, Paolo Giordano, Philipp Petersen, Anna Sztyber-Betley, Paolo Faraboschi, Robin Riblet, Jonathan Crozier, Shiv Halasyamani, Antonella Pinto, Shreyas Verma, Prashant Joshi, Eli Meril, Zheng-Xin Yong, Allison Tee, Jérémy Andréoletti, Orion Weller, Raghav Singhal, Gang Zhang, Alexander Ivanov, Seri Khoury, Nils Gustafsson, Hamid Mostaghimi, Kunvar Thaman, Qijia Chen, Tran Quoc Khánh, Jacob Loader, Stefano Cavalleri, Hannah Szlyk, Zachary Brown, Himanshu Narayan, Jonathan Roberts, William Alley, Kunyang Sun, Ryan Stendall, Max Lamparth, Anka Reuel, Ting Wang, Hanmeng Xu, Pablo Hernández-Cámara, Freddie Martin, Thomas Preu, Tomek Korbak, Marcus Abramovitch, Dominic Williamson, Ida Bosio, Ziye Chen, Biró Bálint, Eve J. Y. Lo, Maria Inês S. Nunes, Yibo Jiang, M Saiful Bari, Peyman Kassani, Zihao Wang, Behzad Ansarinejad, Yewen Sun, Stephane Durand, Guillaume Douville, Daniel Tordera, George Balabanian, Earth Anderson, Lynna Kvistad, Alejandro José Moyano, Hsiaoyun Milliron, Ahmad Sakor, Murat Eron, Isaac C. McAlister, Andrew Favre D. O., Shailesh Shah, Xiaoxiang Zhou, Firuz Kamalov, Ronald Clark, Sherwin Abdoli, Tim Santens, Harrison K Wang, Evan Chen, Alessandro Tomasiello, G. Bruno De Luca, Shi-Zhuo Looi, Vinh-Kha Le, Noam Kolt, Niels Mündler, Avi Semler, Emma Rodman, Jacob Drori, Carl J Fossum, Luk Gloor, Milind Jagota, Ronak Pradeep, Honglu Fan, Tej Shah, Jonathan Eicher, Michael Chen, Kushal Thaman, William Merrill, Moritz Firsching, Carter Harris, Stefan Ciobâcă, Jason Gross, Rohan Pandey, Ilya Gusev, Adam Jones, Shashank Agnihotri, Pavel Zhelnov, Siranut Usawasutsakorn, Mohammadreza Mofayezi, Alexander Piperski, Marc Carauleanu, David K. Zhang, Kostiantyn Dobarskyi, Dylan Ler, Roman Leventov, Ignat Soroko, Thorben Jansen, Scott Creighton, Pascal Lauer, Joshua Duersch, Vage Taamazyan, Dario Bezzi, Wiktor Morak, Wenjie Ma, William Held, Tran Đuc Huy, Ruicheng Xian, Armel Randy Zebaze, Mohanad Mohamed, Julian Noah Leser, Michelle X Yuan, Laila Yacar, Johannes Lengler, Katarzyna Olszewska,

Hossein Shahrtash, Edson Oliveira, Joseph W. Jackson, Daniel Espinosa Gonzalez, Andy Zou, Muthu Chidambaram, Timothy Manik, Hector Haffenden, Dashiell Stander, Ali Dasouqi, Alexander Shen, Emilien Duc, Bita Golshani, David Stap, Mikalai Uzhou, Alina Borisovna Zhidkovskaya, Lukas Lewark, Miguel Orbegozo Rodriguez, Mátyás Vincze, Dustin Wehr, Colin Tang, Shaun Phillips, Fortuna Samuele, Jiang Muzhen, Fredrik Ekström, Angela Hammon, Oam Patel, Faraz Farhidi, George Medley, Forough Mohammadzadeh, Madellene Peñaflor, Haile Kassahun, Alena Friedrich, Claire Sparrow, Rayner Hernandez Perez, Taom Sakal, Omkar Dhamane, Ali Khajegili Mirabadi, Eric Hallman, Kenchi Okutsu, Mike Battaglia, Mohammad Maghsoudimehrabani, Alon Amit, Dave Hulbert, Roberto Pereira, Simon Weber, Handoko, Anton Peristyy, Stephen Malina, Samuel Albanie, Will Cai, Mustafa Mehkary, Rami Aly, Frank Reidegeld, Anna-Katharina Dick, Cary Friday, Jasdeep Sidhu, Hassan Shapourian, Wanyoung Kim, Mariana Costa, Hubeyb Gurdogan, Brian Weber, Harsh Kumar, Tong Jiang, Arunim Agarwal, Chiara Ceconello, Warren S. Vaz, Chao Zhuang, Haon Park, Andrew R. Tawfeek, Daattavya Aggarwal, Michael Kirchhof, Linjie Dai, Evan Kim, Johan Ferret, Yuzhou Wang, Minghao Yan, Krzysztof Burdzy, Lixin Zhang, Antonio Franca, Diana T. Pham, Kang Yong Loh, Joshua Robinson, Abram Jackson, Shreen Gul, Gunjan Chhablani, Zhehang Du, Adrian Cosma, Jesus Colino, Colin White, Jacob Votava, Vladimir Vinnikov, Ethan Delaney, Petr Spelda, Vit Stritecky, Syed M. Shahid, Jean-Christophe Mourrat, Lavr Vetoshkin, Koen Sponselee, Renas Bacho, Florencia de la Rosa, Xiuyu Li, Guillaume Malod, Leon Lang, Julien Laurendeau, Dmitry Kazakov, Fatimah Adesanya, Julien Portier, Lawrence Hollom, Victor Souza, Yuchen Anna Zhou, Julien Degorre, Yiğit Yalın, Gbenga Daniel Obikoya, Luca Arnaboldi, Rai, Filippo Bigi, M. C. Boscá, Oleg Shumar, Kaniuar Bacho, Pierre Clavier, Gabriel Recchia, Mara Popescu, Nikita Shulga, Ngefor Mildred Tanwie, Denis Peskoff, Thomas C. H. Lux, Ben Rank, Colin Ni, Matthew Brooks, Alesia Yakimchyk, Huanxu, Liu, Olle Häggström, Emil Verkama, Hans Gundlach, Leonor Brito-Santana, Brian Amaro, Vivek Vajipey, Rynaa Grover, Yiyang Fan, Gabriel Poesia Reis e Silva, Linwei Xin, Yosi Kratish, Jakub Łucki, Wen-Ding Li, Sivakanth Gopi, Andrea Caciolai, Justin Xu, Kevin Joseph Scaria, Freddie Vargus, Farzad Habibi, Long, Lian, Emanuele Rodolà, Jules Robins, Vincent Cheng, Tony Fruhauff, Brad Raynor, Hao Qi, Xi Jiang, Ben Segev, Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Michael P. Brenner, Mao Mao, Xinyu Zhang, David Avagian, Eshawn Jessica Scipio, Alon Ragoler, Justin Tan, Blake Sims, Rebeka Plecnik, Aaron Kirtland, Omer Faruk Bodur, D. P. Shinde, Zahra Adoul, Mohamed Zekry, Ali Karakoc, Tania C. B. Santos, Samir Shamseldeen, Loukmane Karim, Anna Liakhovitskaia, Nate Resman, Nicholas Farina, Juan Carlos Gonzalez, Gabe Maayan, Sarah Hoback, Rodrigo De Oliveira Pena, Glen Sherman, Elizabeth Kelley, Hodjat Mariji, Rasoul Pouriamanesh, Wentao Wu, Sandra Mendoza, Ismail Alarab, Joshua Cole, Danyelle Ferreira, Bryan Johnson, Mohammad Safdari, Liangti Dai, Siriphan Arthornthurasuk, Alexey Pronin, Jing Fan, Angel Ramirez-Trinidad, Ashley Cartwright, Daphiny Pottmaier, Omid Taheri, David Outevsky, Stanley Stepanic, Samuel Perry, Luke Askew, Raúl Adrián Huerta Rodríguez, Ali M. R. Minissi, Sam Ali, Ricardo Lorena, Krishnamurthy Iyer, Arshad Anil Fasiludeen, Sk Md Salauddin, Murat Islam, Juan Gonzalez, Josh Ducey, Maja Somrak, Vasilios Mavroudis, Eric Vergo, Juehang Qin, Benjámin Borbás, Eric Chu, Jack Lindsey, Anil Radhakrishnan, Antoine Jallon, I. M. J. McInnis, Pawan Kumar, Laxman Prasad Goswami, Daniel Bugas, Nasser Heydari, Ferenc Jeanplong, Archimedes Apronti, Abdallah Galal, Ng Ze-An, Ankit Singh, Joan of Arc Xavier, Kanu Priya Agarwal, Mohammed Berkani, Benedito Alves de Oliveira Junior, Dmitry Malishev, Nicolas Remy, Taylor D. Hartman, Tim Tarver, Stephen Mensah, Javier Gimenez, Roselynn Grace Montecillo, Russell Campbell, Asankhaya Sharma, Khalida Meer, Xavier Alapont, Deepakkumar Patil, Rajat Maheshwari, Abdelkader Dendane, Priti Shukla, Sergei Bogdanov, Sören Möller, Muhammad Rehan Siddiqi, Prajvi Saxena, Himanshu Gupta, Innocent Enyekwe, Ragavendran P V, Zienab EL-Wasif, Aleksandr Maksapetyan, Vivien Rossbach, Chris Harjadi, Mohsen Bahaloohoreh, Song Bian, John Lai, Justine Leon Uro, Greg Bateman, Mohamed Sayed, Ahmed Menshawy, Darling Duclosel, Yashaswini Jain, Ashley Aaron, Murat Tiryakioglu, Sheeshram Siddh, Keith Krenek, Alex Hoover, Joseph McGowan, Tejal Patwardhan, Summer Yue, Alexandr Wang, and Dan Hendrycks. Humanity's last exam, 2025. URL: https://arxiv.org/abs/2501.14249, arXiv:2501.14249.

[6] Jon Saad-Falcon, Adrian Gamarra Lafuente, Shlok Natarajan, Nahum Maru, Hristo Todorov, Etash Guha, E. Kelly Buchanan, Mayee Chen, Neel Guha, Christopher Ré, and Azalia Mirhoseini. Archon: An architecture search framework for inference-time techniques, 2024. URL: https://arxiv.org/abs/2409.15254, arXiv:2409.15254.

[7] Timo Schick, Jane Dwivedi-Yu, Roberta Raileanu Roberto Dessì, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools, 2023. URL: https://arxiv.org/abs/2302.04761, arXiv:2302.04761.

[8] Karthik Valmeekam, Kaya Stechly, and Subbarao Kambhampati. Llms still can't plan; can lrms? a preliminary evaluation of openai's o1 on planbench, 2024. URL: https://arxiv.org/abs/2409.13373, arXiv:2409.13373.

[9] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities, 2024. URL: https://arxiv.org/abs/2406.04692, arXiv:2406.04692.

[10] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. URL: https://arxiv.org/abs/2210.03629, arXiv:2210.03629.

A Appendix

A.1 Custom Component Prompts

A.1.1 Planner

System: “You are an expert at planning and reasoning given a user query. Once you receive a user query, you should do the following steps:

1. Summarize what the user is asking in a clearer way.
2. Provide a step-by-step reasoning process on what you need to do in order to fully answer the user query."

User: “You have been provided with a user’s query: {query}. Please summarize the user query and provide a step-by-step reasoning process to fully answer the user query."

A.1.2 Expander

System: “You are an expert at expanding and enriching responses with additional context and details. When expanding a response:

1. Only add information that is directly relevant to the original response.
2. Maintain factual accuracy and consistency with the original.
3. Focus on adding valuable context that enhances understanding.
4. Keep additions clear and well-structured.
5. Do not contradict or modify the original content."

User: “You have been provided with a user query and an AI’s response to that query. Your task is to enhance this response by providing additional relevant context and details. Keep your additions focused and directly relevant to the original response. User Query: {query} Original Response: {original_response} Please provide additional context, examples, or clarifying details that would enhance the original response. Focus on information that adds value while maintaining relevance to the query. Do not contradict or modify the original response."

A.2 Tool Call Web Search

1. For generating queries
   User: “Given a user’s query and a step-by-step analysis of the solution, output {num_searches} search queries to find more information to help answer the user’s query. Output the search queries in a comma-separated list, do not include any other text. User Query: {query} Step-by-Step Analysis: {reasoning_output}"
2. Augmenting Generation module
   “Use these search results to improve your answer to the user’s query: {search_results"

A.3 Archon Architecture Configurations

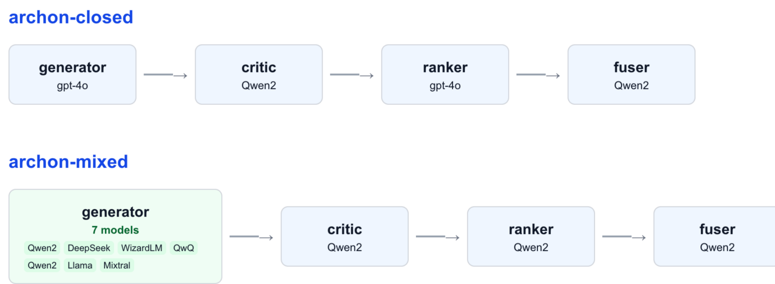

Figure 6: System architecture overview of the Archon framework

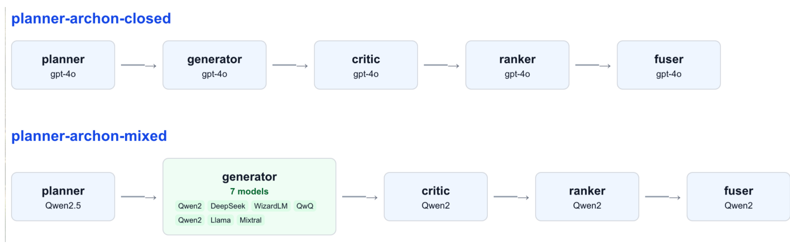

Figure 7: Planner module responsible for decomposing tasks into subtasks

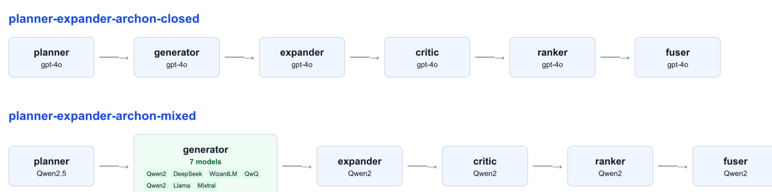

Figure 8: Expander component illustrating task expansion and refinement

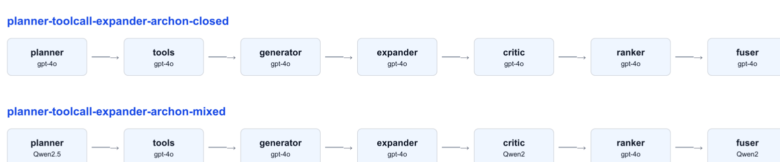

Figure 9: Tool call handler coordinating execution of external tools

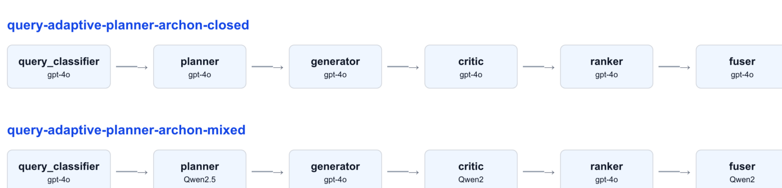

Figure 10: Query adaptive component optimizing prompts based on task context